Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com

Estimating access to drinking water and sanitation: The need to account for uncertainty in trend analysis

F. Ezbakhe *, A. Pérez-Foguet

Department of Civil and Environmental Engineering (DECA), Engineering Sciences and Global Development (EScGD), Barcelona School of Civil Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain

ARTICLE INFO

Article history:
Received 21 June 2019
Received in revised form 6 August 2019
Accepted 6 August 2019
Available online xxx

Editor: Dr. Damia Barcelo

Keywords:
Water, sanitation and hygiene (WASH)
Sampling errors
Household surveys
Compositional data
Joint Monitoring Programme (JMP)
Sustainable development goals

ABSTRACT

Nationally representative household surveys are the main source of data for tracking drinking water, sanitation and hygiene (WASH) coverage. However, all survey point estimates have a certain degree of error that must be considered when interpreting survey results for policy and decision making. In this article, we develop an approach to characterize and quantify uncertainty around WASH estimates. We apply it to four countries – Bolivia, Gambia, Morocco and India – representing different regions, number of data points available and types of trajectories, in order to illustrate the importance of communicating uncertainty for temporal estimates, as well as taking into account both the compositional nature and non-linearity of JMP data. The approach is found to be robust, versatile and particularly useful in the WASH sector, where the dissemination and analysis of standard errors lag behind. While it only considers the uncertainty arising from sampling, the proposed approach can help improve the interpretation of WASH data when evaluating trends in coverage and informing decision making.

1. Introduction

SUBStantial progress has been made worldwide in increasing people's access to water, sanitation and hygiene (WASH) services. Between 1990 and 2015, almost 2.6 billion people gained access to improved drinking water sources and 2.1 billion gained access to improved sanitation facilities (JMP, 2015). Notwithstanding this laudable achievement, there remains a tremendous effort to reach the millions of unserved people (JMP, 2017). This has severe health implications: in 2016 alone, inadequate WASH was estimated to cause 829,000 diarrhoeal deaths that would have been preventable through access to improved WASH services (Prüss-Ustün et al., 2019). Acknowledging the need to further increase access to basic WASH services, the sixth Sustainable Development Goal (SDG 6) of the 2030 Agenda specifically calls for countries to "achieve universal and equitable access to safe and affordable drinking water for all" (target 6.1) as well as "achieve access to adequate and equitable sanitation and hygiene for all and end open defecation" (target 6.2) (UNGA, 2015).

Realizing these ambitious targets will not only require greater investments in WASH, but also understanding the levels and trends in service coverage in order to evaluate countries' progress and identify priorities for improvement (Cronk et al., 2015). The responsibility of monitoring progress of the SDG 6 targets related to WASH lies with the WHO/UNICEF's Joint Monitoring Programme (JMP). Since 1990, the JMP has been producing estimates of national, regional and global progress on WASH access. The JMP currently generates estimates for a total of 26 indicators related to WASH, all of which refer to the proportion of the population using or having a specific water, sanitation or hygiene service level (JMP, 2018). The estimation method used by JMP begins with the identification and compilation of all nationally-representative data relevant to the use of WASH services. A simple linear regression – through ordinary least squares – is then used to model the proportions of the population using each service level over time.

However, the JMP estimation method presents some noteworthy limitations. First, the use of linear regression has been found to introduce substantial bias in estimates, particularly when coverage rates show non-linear patterns (Bartram et al., 2014; Fuller et al., 2016). Furthermore, as Luh et al. (2013) highlight, simple linear regressions fail to capture progressive realization of the human rights to water and sanitation. Several alternative regression approaches have been proposed to address this shortcoming, including quadratic, logit, piecewise linear and generalized additive models (Wolf et al., 2013; JMP, 2014).

Second, the JMP estimation method fails to account for the compositional nature of the data. Proportions of the population are subject to a unit-sum constraint and thus cannot vary independently, which invalidates most standard statistical approaches. Pérez-Foguet et al. (2017) have recently addressed this issue by modelling JMP data with compositional data (CoDa) analysis. They concluded that the log-ratio transformation of data did not only avoid misleading results from when proportions were analysed separately, but also helped im-

* Corresponding author.
Email address: fatine.ezbakhe@upc.edu (F. Ezbakhe)

https://doi.org/10.1016/j.scitotenv.2019.133830
0048-9697/

Table 1
Primary water and sanitation indicators used by the JMP.

	The proportion of the population that uses…
Water	
W_1	All improved drinking water sources
W_2	Piped drinking water sources
W_3	No drinking water sources (i.e., surface water)
Sanitation	
S_1	All improved sanitation facilities
S_2	Improved sanitation facilities connected to sewers
S_3	No sanitation facilities (i.e., open defecation)

prove the performance of regression models, especially when coverage rates were near 0% or 100%.

Third, the characterization and representation of uncertainty around estimates remains an untackled issue by the JMP (JMP, 2014). This is of utmost importance, as estimates are largely based on data from nationally representative household surveys, subject to both sampling and non-sampling errors. Sampling errors arise from the very act of sampling: WASH coverage data are measured from the types of drinking water sources and sanitation/handwashing facilities used in the sampled households, rather than all sources and facilities in the entire population. Measuring the WASH services of another sample of households taken from the same population would give different estimates. Non-sampling errors, on the other hand, arise from biases in data collection, such as duplication or omission of households, inappropriate interview methods, and errors in data processing operations (Banda, 2003). Consequently, in addition to further minimize these errors, uncertainty assessment of the estimates – in the form of confidence intervals for example – is indispensable for an evidence-based analysis of levels and trends in WASH coverage. Failure to conduct and report such confidence intervals may lead to misinterpretation of rates and trends, and ultimately undermine effective policymaking for WASH.

However, reporting confidence intervals in WASH estimates is far from an easy endeavour. First, non-sampling errors are difficult to account for and evaluate statistically (Eisele et al., 2013). Second, information on sampling errors is frequently missing in survey reports and, even where published, it is often unclear whether they have been computed accurately (Betti et al., 2018). Furthermore, the general assumption that estimates from nationally representative household surveys are approximately normally distributed can be problematic when proportions are near zero or one (Janicki, 2019).

As the interest in estimates of WASH coverage will continue to grow in the post-2015 era, we can learn much more by characterizing the uncertainty around WASH estimates. In this paper, we sought to complete the work undertaken by Pérez-Foguet et al. (2017) by examining the uncertainties around water and sanitation estimates. Our aim is to present a fairly simple approach to characterize and communicate uncertainty in WASH trend analysis, while taking into account both the compositional and non-linear nature of the data. In particular, our method tackles two central issues in the WASH sector:

- How can we portray the accuracy of WASH estimates when, unfortunately, information on sampling errors is seldom included in household survey reports?
- How can we model the sampling distribution of WASH estimates near the boundary (i.e., coverage rates of 0% or 100%) where assumptions of normality are no longer valid?

To exemplify our method, we use four case studies – Bolivia, Gambia, Morocco and India –, each representing a different region, number of data points available and type of trend trajectory. Our analysis is structured in three parts. First, we demonstrate the importance of accounting for the compositional nature of WASH data by comparing estimates obtained with standard and compositional approaches. Second, we examine the effect of non-linearity in the data by evaluating the discrepancy between conventional ordinary least squares (OLS) and generalized additive models (GAM). Finally, we assess the magnitude of standard errors and confidence intervals for the WASH estimates.

The rest of the article is organized as follows. We first provide a background on CoDa analysis in Section 2. In Section 3, we describe the JMP database (Section 3.1), the method we propose for characterizing uncertainty in WASH estimates (Section 3.2) and the four case studies selected (Section 3.3). Section 4 presents the results of applying the proposed method. In Section 5, we discuss the main implications of our method for monitoring WASH coverage. Finally, Section 6 highlights the main conclusions of the article.

2. Background on coda analysis

Compositional data are arrays of non-negative multivariate data in which the components represent some part of a whole. There are usually recorded in closed form, summing to a constant (e.g. proportions summing to 1 or percentages summing to 100%). Such data are widespread in many disciplines, such as geosciences, biology, economics

Table 2
Data availability for each indicator included in the JMP database (1990–2015).

Service	Setting	Indicator	Number of countries with the following number of data points						
			0	1	2	3–5	6–10	11–15	>16
Water	Urban	W_1. Improved	29	13	17	46	48	28	48
		W_2. Piped	31	15	18	43	47	28	47
		W_3. Surface	93	16	10	25	34	19	32
		W_{123}. All indicators	93	16	10	25	36	18	31
Water	Rural	W_1. Improved	34	16	15	42	46	32	44
		W_2. Piped	36	19	14	40	46	31	43
		W_3. Surface	93	16	10	25	33	19	33
		W_{123}. All indicators	93	16	10	25	36	17	32
Sanitation	Urban	S_1. Improved	23	23	10	35	57	26	55
		S_2. Sewer	32	25	19	49	48	20	36
		S_3. Open defecation	80	17	12	34	35	20	31
		S_{123}. All indicators	87	20	17	36	29	19	21
Sanitation	Rural	S_1. Improved	27	23	9	33	56	27	54
		S_2. Sewer	42	24	14	46	47	23	33
		S_3. Open defecation	82	17	12	32	34	21	31
		S_{123}. All indicators	88	21	16	35	29	18	22

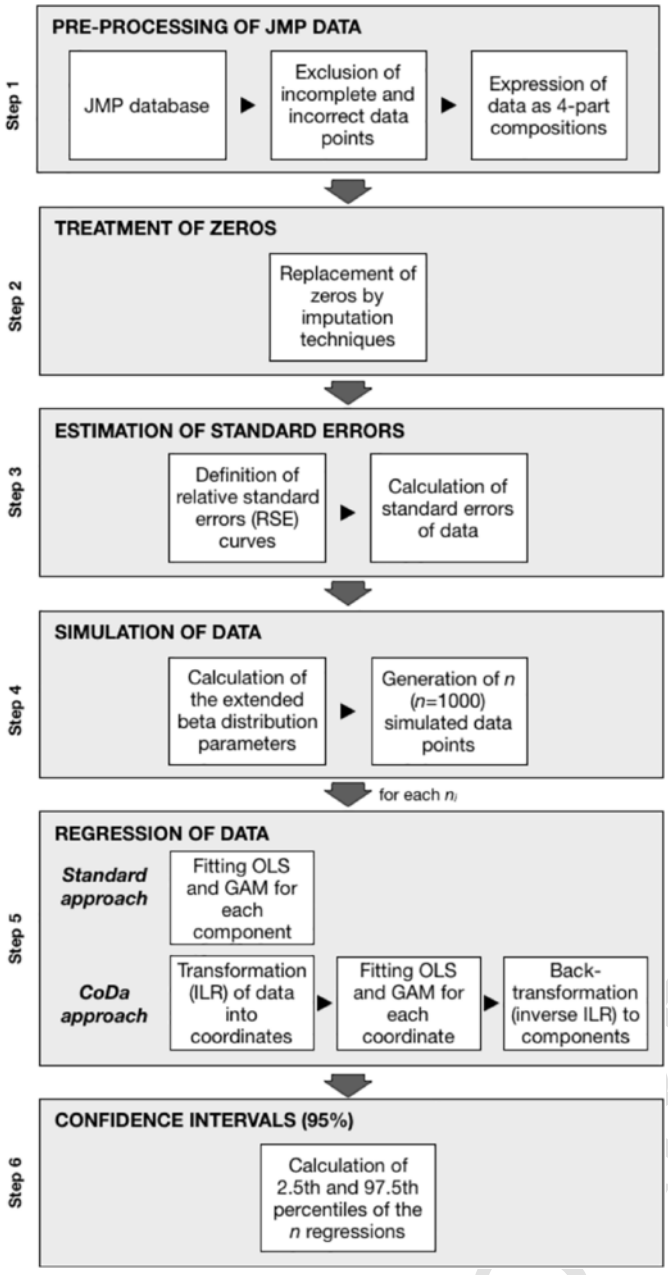

Fig. 1. Overview of the proposed approach for characterizing uncertainty in WASH estimates while considering the compositional and non-linear nature of data.

and population studies (Lloyd et al., 2012; Ferrer-Rosell et al., 2016; Bergeron-Boucher et al., 2017; Wei et al., 2018; Linares-Mustaros et al., 2018; Marcillo-Delgado et al., 2019). By definition, JMP data are compositional: the individual proportions of the population using each WASH service level are not independent of each other, but related by being expressed as percentages of the total population.

Table 3
4-part compositions considered for this study.

	Means of estimation	The proportion of the population that uses…
Water		
x_1	W_2	Piped drinking water sources
x_2	$W_1 - W_2$	Other improved drinking water sources
x_3	W_3	No drinking water facility (surface water)
x_4	$1 - W_1 - W_3$	Other unimproved drinking water sources
Sanitation		
x_1	S_2	Improved sanitation facilities connected to sewers
x_2	$S_1 - S_2$	Other improved sanitation facilities
x_3	S_3	No sanitation facilities (open defecation)
x_4	$1 - S_1 - S_3$	Other unimproved sanitation facilities

Compositional data have particular and essential properties that arise from the fact that they represent parts of some whole (Pawlowsky-Glahn and Egozcue, 2006). They are a vector of strictly positive real numbers with a constant sum constraint:

$$x = (x_1, \ldots, x_D);\, x_i > 0, i = 1, 2, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \quad (1)$$

The elements of a composition, x_i, are called components or parts, and the only relevant information is contained in the ratios between components (Pawlowsky-Glahn et al., 2015). This conditions the relationships that variables have to one another, and manifests in their variance-covariance structure (Aitchison, 1982). Furthermore, it means that compositional data are enclosed in a subspace where they can only vary between 0 and the radix value. Such subspace – known as the simplex – does not follow the rules of Euclidean geometry, making all standard techniques devised for unconstrained data inappropriate for the analysis of compositional data (Aitchison, 1986, 1999). However, because of its geometry, the simplex can be difficult to work in. As an alternative, the compositional data may be transformed to the real scale where classical statistical procedures can be applied (Pawlowsky-Glahn et al., 2015). These transformations are based on log-ratios between components, and lead to "open" data – called coordinates – that can take any real value between $-\infty$ and $+\infty$. Several log-transformation approaches have been proposed, including the additive log-ratio (ALR), the centred log-ratio (CLR) and the isometric log-ratio (ILR) (Aitchison, 1986; Egozcue et al., 2003).

In the following, ILR transformation is applied to perform the statistical analysis of JMP data. This transformation represents the composition given a particular orthonormal basis in the simplex (Egozcue et al., 2003), given by:

Table 4
Case studies included in the study (x_1, x_2, x_3, x_4 refer to the 4-parts compositions described in Table 3). [Note: trajectories are characterized following the method suggested by Fuller et al., 2016.]

Country	Service	Setting	Region	Data points	Trajectories			
					x_1	x_2	x_3	x_4
Bolivia	Water	Urban	Latin America and the Caribbean	25	Saturation	Acceleration	No change	Linear decline
Gambia	Water	Urban	Sub-Saharan Africa	7	Linear growth	Linear decline	Deceleration	Linear decline
Morocco	Sanitation	Rural	Northern Africa and Western Asia	8	Linear decline	Linear growth	Linear decline	No change
India	Sanitation	Rural	Central and Southern Asia	12	Linear growth	Linear growth	Negative acceleration	No change

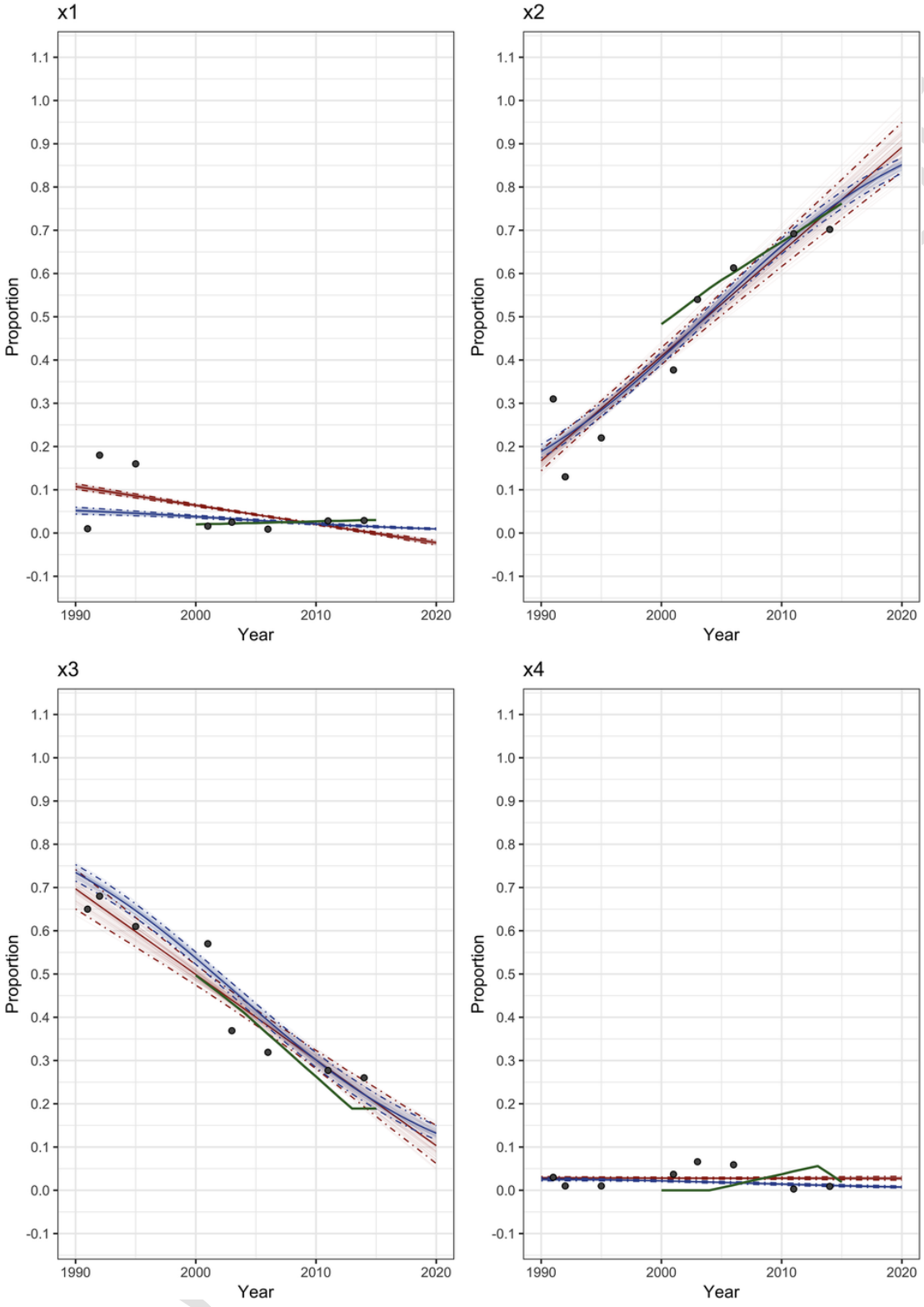

Fig. 2. Coverage estimates for rural sanitation in Morocco, with OLS regression. [Note 1: in red, estimates with the standard approach; in blue, estimates with the compositional approach; in green, estimates provided by the JMP] [Note 2: Estimates are provided with 95% confidence intervals] [Note 3: x1, x2, x3 and x4 refer to the proportion of the population using sewer connections, other improved sanitation, open defecation and other unimproved sanitation, respectively]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

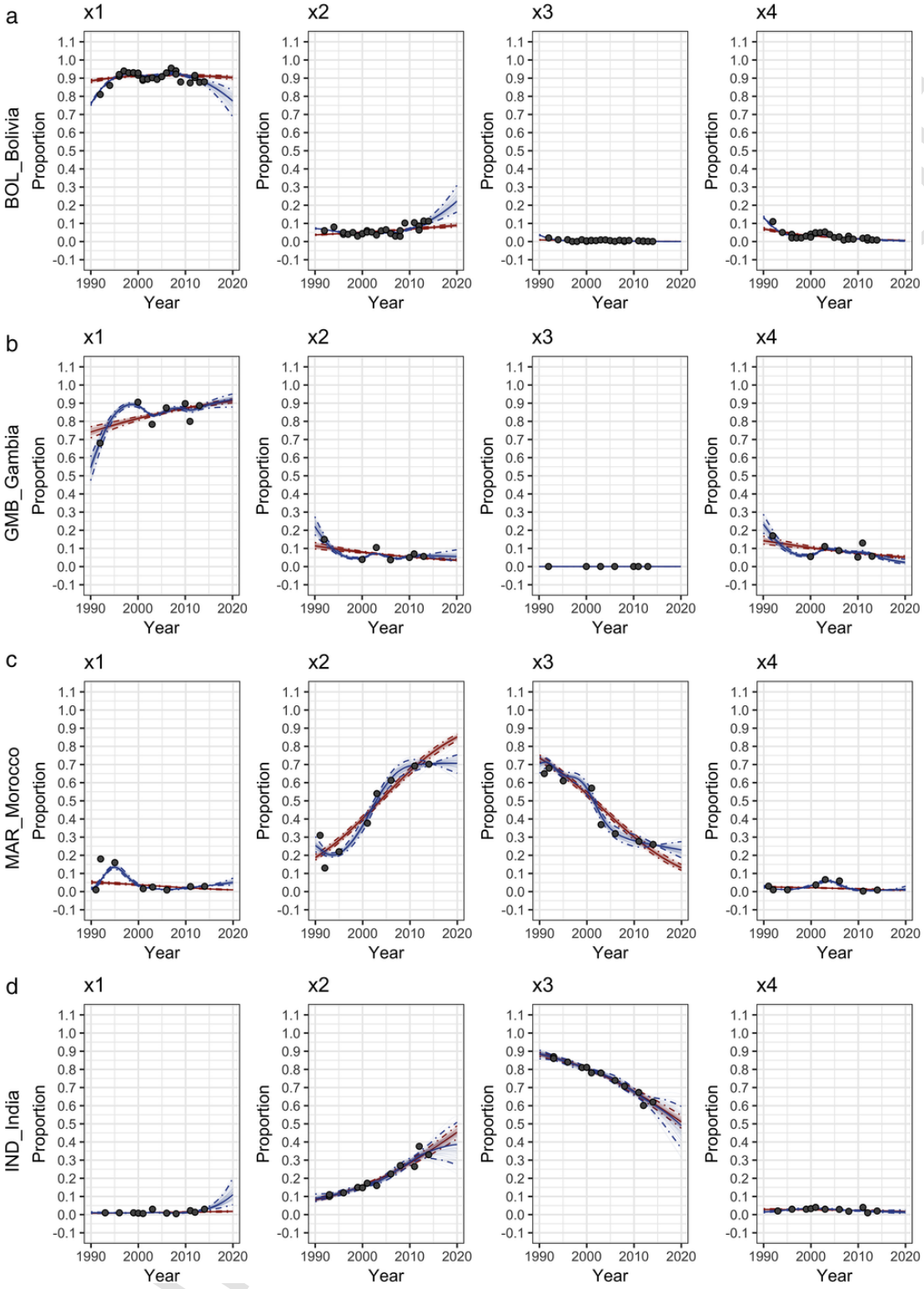

Fig. 3. Coverage estimates for (a) urban water in Bolivia, (b) urban water in Gambia, (c) rural sanitation in Morocco and (d) rural sanitation in India, with the compositional approach. [Note 1: in red, estimates from OLS regression; in blue, estimates from GAM regression] [Note 2: Estimates are provided with 95% confidence intervals] [Note 3: x1, x2, x3 and x4 refer to the proportion of the population using piped/sewer connections, other improved water/sanitation, surface water/open defecation and other unimproved water/sanitation, respectively] (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Values of the root-mean-square error (RMSE) and the Nash-Sutcliffe Efficiency coefficient (NSE) for coverage estimates with OLS and GAM regression models. [Note 1: estimates are obtained with the compositional approach.] [Note 2: For water, x_1, x_2, x_3 and x_4 refer to the proportion of the population using piped connections, other improved water, surface water and other unimproved water, respectively; for sanitation, they indicate sewer connections, other improved sanitation, open defecation and other unimproved sanitation.]

Case study	Component	RSME		NSE	
		OLS	GAM	OLS	GAM
Urban water in Bolivia	x_1	0.0288	0.0183	0.0930	0.6315
	x_2	0.0211	0.0155	0.2894	0.6154
	x_3	0.0040	0.0031	0.2046	0.5431
	x_4	0.0161	0.0102	0.4437	0.7779
Urban water in Gambia	x_1	0.0584	0.0337	0.4183	0.8062
	x_2	0.0287	0.0147	0.4344	0.8527
	x_3	0.0000	0.0000	NA[a]	NA[a]
	x_4	0.0337	0.0228	0.3390	0.6970
Rural sanitation in Morocco	x_1	0.0640	0.0437	0.0514	0.5576
	x_2	0.0660	0.0435	0.8967	0.9551
	x_3	0.0568	0.0360	0.8827	0.9528
	x_4	0.0242	0.0086	−0.1327	0.8568
Rural sanitation in India	x_1	0.0078	0.0061	0.1483	0.4701
	x_2	0.0235	0.0215	0.9262	0.9378
	x_3	0.0150	0.0136	0.9702	0.9753
	x_4	0.0086	0.0068	0.0298	0.4010

[a] Note: NA values are obtained because all observed data are equal.

$$\boldsymbol{y} = ilr\ (\boldsymbol{x}) = \log(\boldsymbol{x}) \bullet \boldsymbol{V} \tag{2}$$

where $\boldsymbol{x}$ is the vector with the D parts of the composition, $\boldsymbol{V}$ a D x $(D-1)$ matrix representing the orthonormal basis in the simplex, and $\boldsymbol{y}$ the resulting vector with the $D-1$ coordinates of the composition in that basis $\boldsymbol{V}$.

There are several ways to define orthonormal bases in the simplex, one of which consists in a sequential binary partition (SBP) of the composition (Pawlowsky-Glahn et al., 2015). A SBP represents a hierarchy of the parts of a composition, and contains successive splits of the parts into two groups, coded by the signs + and − respectively (Pawlowsky-Glahn and Egozcue, 2011). The orthonormal basis $\boldsymbol{V}$ can be obtained from the SBP as:

$$y_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \bullet \log\left(\frac{\left(\prod_{+} x_{ij}\right)^{\frac{1}{r_i}}}{\left(\prod_{-} x_{ik}\right)^{\frac{1}{s_i}}}\right); i = 1, \ldots, D-1 \tag{3}$$

where y_i is the i[th] coordinate (or balance) of the composition, x_{ij} and x_{ik} are the components coded as + and − in the i[th] partition, and r_i and s_i are the number of components coded as + and −, respectively.

Once data are transformed into ILR balances, standard statistical approaches can be applied. Finally, regression points can be back-transformed to the original space using the inverse ILR:

$$\boldsymbol{x} = \mathscr{C}\ \left[\exp(\boldsymbol{V}\boldsymbol{y})\right] \tag{4}$$

where $\boldsymbol{y}$ contains the ILR coordinates of $\boldsymbol{x}$ with respect to the basis $\boldsymbol{V}$,

Table 6
Coverage estimates with OLS and GAM regression models for years 1990, 2000, 2010 and 2020. [Note 1: Values are given in percentages of the population.] [Note 2: In parenthesis the 95% confidence intervals.] [Note 3: For water, x_1, x_2, x_3 and x_4 refer to the proportion of the population using piped connections, other improved water, surface water and other unimproved water, respectively; for sanitation, they indicate sewer connections, other improved sanitation, open defecation and other unimproved sanitation.]

Case study	Component	OLS				GAM			
		1990	2000	2010	2020	1990	2000	2010	2020
Urban water in Bolivia	x_1	88.4 (87.5–89.3)	91.3 (90.9–91.6)	91.5 (91.0–92.0)	90.3 (89.2–91.3)	75.6 (74.8–76.5)	91.9 (91.4–92.3)	91.5 (90.8–92.1)	77 (67.6–83.4)
	x_2	3.7 (3.4–4.0)	5.1 (4.8–5.3)	6.8 (6.4–7.2)	9.0 (8.0–10.0)	7.2 (6.9–7.5)	4.5 (4.2–4.8)	6.9 (6.3–7.5)	22.6 (16.2–31.8)
	x_3	0.9 (0.8–1.1)	0.4 (0.4–0.5)	0.2 (0.2–0.2)	0.1 (0.1–0.1)	3.7 (3.4–4.0)	0.3 (0.3–0.4)	0.2 (0.2–0.3)	0.0 (0.0–0.0)
	x_4	7.0 (6.2–7.8)	3.3 (3.1–3.4)	1.5 (1.4–1.6)	0.7 (0.6–0.8)	13.5 (12.9–14.2)	3.3 (3.1–3.6)	1.4 (1.3–1.6)	0.4 (0.2–0.7)
Urban water in Gambia	x_1	74.1 (70.8–77.1)	81.6 (80.4–82.7)	87.2 (86.5–87.9)	91.3 (90.1–92.4)	54.5 (47.5–60.8)	88.7 (87.7–89.6)	86.8 (85.5–87.8)	91.9 (87.5–94.8)
	x_2	11.4 (9.7–13.2)	7.9 (7.3–8.5)	5.4 (5.0–5.8)	3.6 (3–4.2)	22.1 (17.6–27.1)	4.9 (4.4–5.5)	5.4 (4.8–6.0)	5.7 (3.2–9.3)
	x_3	0.1 (0.0–0.1)	0.1 (0.0–0.1)	0.1 (0.0–0.1)	0.0 (0.0–0.1)	0.0 (0.0–0.1)	0.1 (0.0–0.1)	0.1 (0.0–0.1)	0.0 (0.0–0.1)
	x_4	14.4 (12.3–16.7)	10.4 (9.6–11.2)	7.3 (6.8–7.9)	5.1 (4.3–5.9)	23.4 (18.4–28.7)	6.4 (5.7–7.1)	7.8 (7.1–8.6)	2.4 (1.4–4.0)
Rural sanitation in Morocco	x_1	5.2 (4.4–6.0)	3.8 (3.5–4.2)	2.1 (1.9–2.4)	0.9 (0.8–1.2)	1.6 (1.1–2.2)	3.6 (2.9–4.4)	2 (1.5–2.4)	5.1 (3.7–7.2)
	x_2	18.8 (17.2–20.4)	40.3 (38.9–41.8)	66.3 (64.5–68.2)	85.1 (83.3–86.8)	25.4 (21.4–29.7)	36.3 (33.1–39.6)	69.1 (65.7–72.6)	70.5 (65.0–75.5)
	x_3	73.5 (71.5–75.3)	53.7 (52.2–55.3)	30.1 (28.3–31.9)	13.2 (11.6–14.9)	70.4 (65.7–74.4)	56.7 (53.3–60)	28.1 (24.7–31.5)	22.9 (18.6–28.2)
	x_4	2.5 (2.2–3.0)	2.2 (2.0–2.4)	1.4 (1.2–1.7)	0.7 (0.5–1.0)	2.7 (2.0–3.5)	3.3 (2.8–3.9)	0.8 (0.5–1.1)	1.5 (0.7–2.9)
Rural sanitation in India	x_1	0.7 (0.6–0.9)	1 (0.9–1.2)	1.4 (1.3–1.6)	1.8 (1.4–2.2)	1.1 (0.7–1.6)	0.9 (0.8–1.1)	1.1 (0.9–1.4)	11.5 (5.6–19.9)
	x_2	8.2 (7.3–9.2)	15.8 (15–16.6)	28.4 (26.8–30)	45.4 (41.7–49.3)	9 (7.2–11.4)	15.3 (14.4–16.2)	29.1 (27.2–31.1)	38.4 (27.1–50.6)
	x_3	88.3 (87–89.4)	80.5 (79.7–81.4)	67.9 (66.1–69.6)	51.1 (47.2–54.8)	88.5 (85.7–90.5)	80.4 (79.3–81.4)	67.6 (65.6–69.6)	48.7 (36.0–61.0)
	x_4	2.8 (2.4–3.3)	2.6 (2.5–2.8)	2.3 (2.0–2.6)	1.8 (1.4–2.2)	1.5 (1.1–2.0)	3.4 (3.1–3.7)	2.2 (1.9–2.5)	1.4 (0.6–2.6)

Table 7
Minimum and maximum RSE values considered in the sensibility analysis.

Set	RSE	Type of data source		
		MICS,DHS,LSMS,WHS	Other household surveys	Censuses
A	Min	4%	8%	0%
	Max	40%	80%	0%
B	Min	2%	20%	0%
	Max	4%	40%	0%
C	Min	1%	1.5%	0%
	Max	10%	15%	0%

and $\mathscr{C}$ is the closure operator:

$$\mathscr{C}[\boldsymbol{x}] = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right) \quad (5)$$

For a 4-part composition, $\boldsymbol{x}=(x_1, x_2, x_3, x_4)$, an example SBP can be:

Order	x_1	x_2	x_3	x_4	r	s
1	+1	+1	−1	−1	2	2
2	+1	−1	0	0	1	1

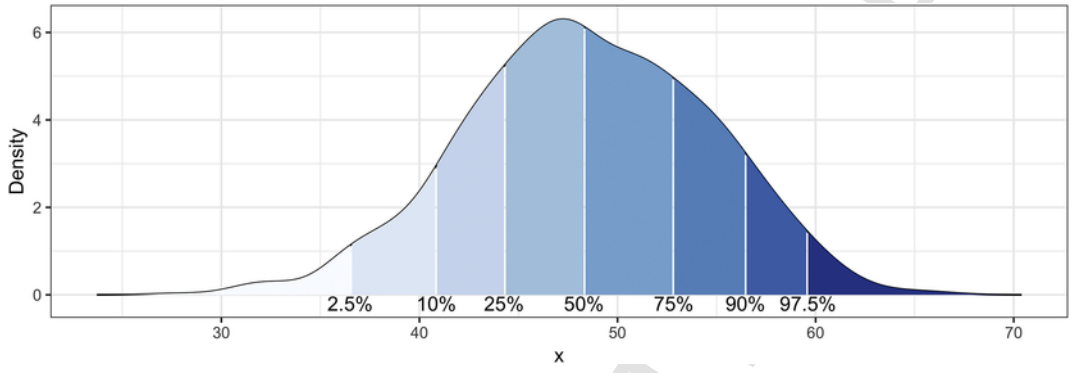

Fig. 4. Sampling distribution of the percentage of the population using open defecation (x_3) in rural India in 2020.

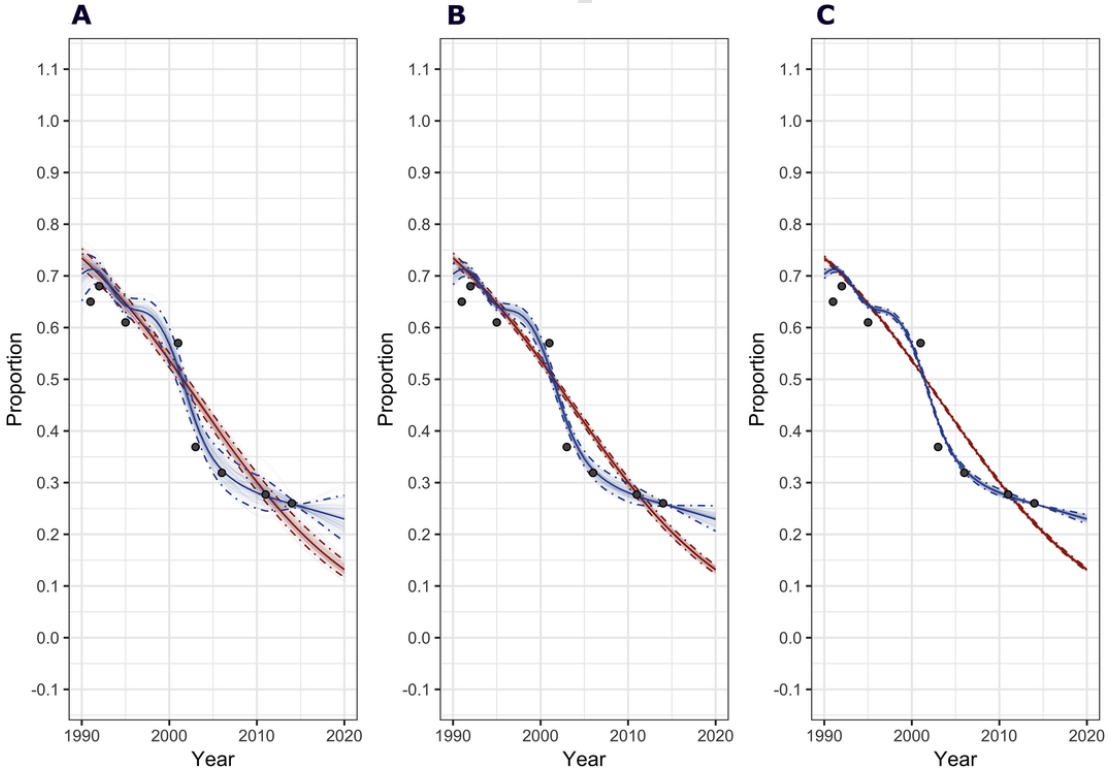

Fig. 5. Coverage estimates for access to open defecation (x_3) in rural Morocco, with the compositional approach. [Note 1: in red, estimates from OLS regression; in blue, estimates from GAM regression.] [Note 2: Estimates are provided with 95% confidence intervals.] [Note 3: A, B and C refer to the different RSE curves from Table 7.] (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3	0	0	+1	−1	1	1

and therefore, the orthonormal basis:

$$V = \begin{bmatrix} & & 1/\sqrt{2} & 0 \\ 1/21/2 & -1/2 & -1/\sqrt{2} & 0 \\ & -1/2 & 0 & 1/\sqrt{2} \\ & & 0 & -1/\sqrt{2} \end{bmatrix}$$

The ILR coordinates can be computed, following Eq. (3), as:

$$y_1 = \frac{1}{2}\log\frac{x_1 x_2}{x_3 x_4} \tag{4}$$

$$y_2 = \frac{1}{\sqrt{2}}\log\frac{x_1}{x_2} \tag{5}$$

$$y_3 = \frac{1}{\sqrt{2}}\log\frac{x_3}{x_4} \tag{6}$$

3. Methodology

3.1. JMP database

The JMP currently monitors coverage of WASH services in 230 countries and territories, and maintains individual files for each, publicly available at the JMP household database (https://washdata.org/data, data extracted on May 29th, 2019). Six primary indicators are used to monitor water and sanitation access, each reported separately for urban and rural populations (Table 1).

JMP data are given in percentages of the population with a precision of 1 digit after the decimal point (i.e., 3 digits for proportions of the population). For a given indicator, the data points available vary depending on the country (Table 2). More than one third of all countries lack any data on surface water and open defecation (34.9%–40.6%), and nearly four fifths have <10 data points for these two indicators (77.3%–77.7%). Furthermore, most countries do not provide complete data for all three indicators: only 21.4% and 17.5% of all countries present >10 complete data points for water and sanitation, respectively.

JMP data are obtained from nationally representative household surveys – including Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), World Health Surveys (WHS) and Living Standards Measurement Study (LSMS) – and national censuses conducted by governments. Since survey data are based on household samples, they will differ somewhat from data that would have been obtained from a complete census. This sampling error is measured by the standard error statistic, which reflects the variability between estimates we would obtain from different samples of the population.

However, the JMP does not report the standard errors as part of its household database. As Bain et al. (2018) state, this is partly due to "concerns that non-sampling errors are likely to dominate sampling errors, especially since the underlying household survey data used to assess basic services often have large sample sizes". While it is true that sampling errors represent only one component of the total survey error and may underestimate non-sampling errors, they still need to be accounted for. For instance, in Burkina Faso, the relative standard errors (i.e., the standard error expressed as a fraction of the estimated proportion) of water and sanitation estimates range between 0.7% and 19.2%, with an average of 7.1% (WHS, 2003; MICS, 2006). The challenge, however, lies in obtaining all standard errors of JMP data, as they are not available in the majority of survey reports. For instance, in DHS reports, errors are only provided for a small selection of variables that does not include those water and sanitation-related (Verma and Lê, 1996; Vaessen et al., 2005). Although this complicates the assessment, it is still necessary to gauge the standard errors to characterize the uncertainty around WASH estimates.

3.2. Proposed approach

To characterize uncertainty around WASH estimates while taking into considering both its compositional and non-linear nature, our approach encompasses the following steps: (i) pre-process JMP data in order to express them as 4-part compositions, (ii) treat the zero values in the compositional data by imputation techniques, (iii) estimate the standard errors of the proportions by the use of a generalized relative standard error function, (iv) generate *n* simulations of the compositions for each year following an extended Beta distribution, (v) fit the regression model to the data, and (vi) calculate the 95% confidence intervals from the regression percentiles.

The approach, illustrated in Fig. 1, is undertaken with the software R, using packages *PearsonDS* (Becker and Klößner, 2017), *Compositions* (Gerald van den Boogaart et al., 2018) and *Gam* (Hastie, 2018). The R script can be found in Ezbakhe and Pérez-Foguet (2019).

3.2.1. Step 1. Pre-processing of JMP data

The 3 primary indicators for water (i.e., W_1, W_2 and W_3) and sanitation (S_1, S_2 and S_3) included in the JMP database are analysed as 4-part compositions, as shown in Table 3.

Therefore, only years with complete data for all 3 indicators (i.e., all parts of the composition) are included in the analysis. Furthermore, years with out-of-range data (i.e., $W_1+W_3>1$ and $W_1<W_2$ for water, and $S_1+S_3>1$ and $S_1<S_2$ for sanitation) are excluded. For instance, according to the JMP database, the percentages of people using the different types of drinking water sources in urban Botswana in 2007 were 98.9% (W_1), 99.0% (W_2) and 1.3% (W_3), which is visibly erroneous: the sum of the people using improved and unimproved cannot exceed 100%, and, since piped water is one of the several sources qualified as improved, the people using piped supplies cannot be greater than those using all forms of improved sources.

3.2.2. Step 2. Treatment of zeros

For the compositional analysis of JMP data – based on log-ratios of parts – to be possible, zeros must be first treated. In this case, it is conceptually sound to consider zeros as non-structural zeros, since data are mainly sourced from household surveys. In order words, since we cannot be sure that households using a particular WASH service level do not exist, zeros can be seen as rounded zeros. In such case, zeros can be replaced with Martín-Fernández et al. (2003) imputation technique:

$$r_j = \begin{cases} \delta_j, if\, x_j = 0 \\ \left(1 - \frac{\Sigma_{k|x_j=0}\delta_j}{c}\right) x_j \; if\, x_j > 0 \end{cases} \tag{7}$$

where r_j is the non-zero composition, δ_j is the imputed value on the part x_j, and c is the constant sum-constraint (i.e., $c=1$). In this study, δ_j is associated to the precision of the data, which, as indicated in Section 3, are given with 3 digits. Thus, $\delta=0.5\bullet10^{-3}$.

3.2.3. Step 3. Estimation of standard errors

To overcome the problem of non-reporting of sampling errors of survey data, we use a generalized relative standard error function, which defines a relationship between the relative standard errors (i.e., the standard errors expressed as a percentage of the estimated proportion) and the corresponding proportion. We use a modified version of Gabrel and Jones (2000) formula:

$$RSE = \sqrt{a + b\frac{1-p}{p}} \tag{8}$$

where p is the estimated proportion (i.e., x_i from our compositional data), and a and b are coefficients derived from RSE curves. These RSE curves indicate the magnitude of the relative standard error for estimated proportions of various sizes and should be interpreted as an approximation rather than exact values for any specific proportion. They have the following meaning: for small values of p the relative standard errors are relatively high (when p approaches zero, the relative standard error approaches infinite), and decrease in a square root way as p increases, reaching its minimum value for $p=1$. Therefore, coefficients a and b can be obtained by fixing a minimum and maximum relative standard errors:

$$a = RSE_{\min}{}^2 \tag{9}$$

$$b = \frac{\delta}{1-\delta}\left(RSE_{\max}{}^2 - RSE_{\min}{}^2\right) \tag{10}$$

where RSE_{min} and RSE_{max} are the minimum and maximum relative standard errors, and δ is the precision of the data.

To fix RSE_{min} and RSE_{max}, we distinguish between three types of sources:

(i) In household surveys from MICS, DHS, LSMS and WHS, that are considered to have higher quality, the minimum and maximum relative standard errors are set as 4% and 40%, respectively.
(ii) In other households surveys, they are set as 8% and 80%.
(iii) In censuses, where all households are counted, they are considered zero.

The standard errors are calculated by multiplying the RSE by the estimated proportion. For example, for a proportion of 0.0005, the standard error would be 0.0002 or 0.0004 depending on the type of household survey, while for a proportion of 1 the standard error would equal 0.04 or 0.08. A sensibility analysis will be included to assess the effect of the RSE_{min} and RSE_{max} values on the resulting confidence intervals.

3.2.4. Step 4. Simulation of data

Confidence intervals of estimates are constructed via simulation techniques. This involves generating n simulations (in this case $n=1000$) of the compositions for each year assuming a generalized beta distribution, also known as Pearson Type I (Bowman and Shenton, 2007). The use of a generalized beta distribution – instead of the normal distribution – is motivated by its ability to model proportions near the boundaries (i.e., 0 or 1), where the normal approximation of the sampling distribution is no longer valid (Cameron, 2011). Essentially, Pearson Type I distributions are location-scale transformations of Beta distributions. Its probability density function, for $l \le x \le u$, and shape parameters $\alpha, \beta > 0$, is a power function of the variable x and its reflection as follows:

$$B(x; l, u, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(x-l)^{\alpha-1}(u-x)^{\beta-1}}{(u-l)^{\alpha-\beta-1}} \tag{11}$$

where $\Gamma(k)$ is the complete gamma function.

We use the method of moments estimators to obtain the parameters α and β of the generalized beta distributions (Bain and Engelhardt, 1991). This involves equating the moments of the generalized beta distribution with the sample proportion and variance:

$$p = l + (u-l)\frac{\alpha}{\alpha+\beta} \tag{12}$$

$$se^2 = (u-l)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{13}$$

We express the limits l and u in terms of the precision of the data (i.e., $l=0.5\bullet10^{-3}$ and $u=9.5\bullet10^{-3}$). With these l, u, α, β parameters, we simulate 1000 random proportions using the generalized beta probability distribution.

3.2.5. Step 5. Regressions of data

In each simulation, a regression model is fitted to the data. To evaluate the compositional nature of the data, two statistical approaches are employed: (i) standard approach, where the 4 components are modelled separately, as in the JMP estimation method; and (ii) compositional approach, in which the 4 components are first log-transformed into 3 coordinates, which are then modelled separately and finally regression results are back-transformed, as explained in Section 2.

Furthermore, to analyse the effect of non-linear patterns on estimates, two regression models are applied: (i) ordinary least squares (OLS) linear regression, which is the method used by the JMP; and (ii) standard generalized additive model (GAM), in which the linear form is replaced by a sum of smooth functions. In this case, GAM procedure is applied with thin-plate regression splines with four degrees of freedom (Wood, 2003).

To compare OLS and GAM regression models, we use the root-mean-square error (RMSE) and Nash-Sutcliffe Efficiency (NSE) coefficients, as done by Pérez-Foguet et al. (2017). The RMSE represents the quadratic mean of the differences between the observed and the modelled proportions, whereas the NSE also computes the square differences between the observed values and their mean. In RMSE, a coefficient of 0 would indicate a perfect fit to the data. In NSE, an efficiency of 1 corresponds to a perfect match model and observations, while a value of 0 indicates that the model performs equally to the mean of the observed data and values lower than 0 occur when the observed mean is a better predictor than the model.

3.2.6. Step 6. Confidence intervals

Finally, the 95% confidence intervals are calculated from the 2.5th and 97.5th percentiles of the regressions results.

3.3. Case studies

Although the proposed approach can be applied to any country or territory included in the JMP database, in this article we illustrate it with 4 case studies: (i) urban water in Bolivia; (ii) urban water in Gambia; (iii) rural sanitation in Morocco; and (iv) rural sanitation in India. These four case studies are selected to represent different SDG regions, number of data points and types of trajectories (Table 4).

4. Results

This section presents the results from applying our proposed method to the four case studies. First, we illustrate the importance of considering the compositional nature of WASH data by comparing the results obtained with the standard and compositional approaches. Then, we compare the estimates from OLS and GAM regression models to evaluate the effect of non-linear patterns in the data. Finally, we quantify the confidence intervals of water and sanitation estimates and analyse the effect of the relative standard errors on these confidence intervals.

4.1. The compositional nature of WASH data

Fig. 2 shows the coverage estimates for the case of rural sanitation in Morocco, obtained with the standard statistical approach (i.e., OLS regression model fitted to each indicator separately) and our compositional approach. It also shows (in green) the official estimates provided by the JMP.

According to our standard estimates, the percentages of the population using each service level in 2015 are: −0.1% for sewer connections (x1), 77.1% for other improved facilities (x2), 20.2% for open defecation (x3) and 2.8% for other unimproved facilities (x4). These figures differ slightly from those estimated by the JMP (3%, 76.2%, 18.% and 1.9%, respectively) for two main reasons. First, our estimates are constructed from all data points available, whereas JMP only uses data from 2000 onwards. This has an important effect on the trend of service coverage. For sewer connections (x1), for example, excluding the data points before 2000 leads to a growth trajectory instead of a decline. Second, although by definition OLS regression estimates would add up to 100%, there is no way to ensure that all four values lie between 0 and 1. The current JMP method avoids out-of-boundary values by adjusting the extrapolation results: if estimates are below 0% or above 100%, they are fixed at 0% and 100%, respectively (JMP, 2018).

However, this ad hoc post-process does not address the underlying problem: WASH data are compositional, and ignoring their compositional nature would lead to erroneous results. Consequently, a compositional approach must be applied to obtain more theoretically sound estimates of the different proportions of the population using or having access to water or hygiene facilities, especially when coverage rates are near 0% and 100%.

4.2. Non-linear patterns in WASH data

Fig. 3.3 illustrates the coverage estimates for all case studies, obtained with linear (OLS) and non-linear (GAM) regression analyses. For countries with linear trajectories in their water and sanitation coverage, the differences between OLS and GAM regression models are not substantial. For instance, in India (Fig. 3.d), where three out of the four components present linear patters (i.e., x1, x2 and x4), OLS and GAM estimates never differ by >2 percentage points. However, when trajectories are non-linear, OLS regression tends to under or overestimate coverage levels. This is evident in the case of Bolivia. For piped water (x1), where data points follow a saturation trajectory, OLS overvalues coverage in the 1990–1995 and 2011–2015 periods, but undervalues the coverage levels between 1995 and 2011. The reverse occurs for other improved water sources (x2), in which progress shows an accelerated pattern.

In order to further compare OLS and GAM regression estimates, we use the RMSE and NSE coefficients to quantitatively describe the accuracy of the models to the JMP data (Table 5). In all case studies, GAM provides better RMSE and NSE values, which translates in an improved accuracy of the regression models. This is particularly noticeable in non-linear trends. In Bolivia, the access to other improved water sources (x2), with an acceleration pattern, presents a 0.0056 reduction in the root-mean-square errors; compared to the 0.0009 decrease in the case of access to surface water (x3), where the trajectory is linear.

Despite the "superior" statistical power of non-linear regression models such as GAM, the JMP still choose OLS for the estimation of coverage. There are two main reasons for this. First, the majority of countries show linear trajectories (62%–85.3%, depending on the service and setting (Fuller et al., 2016)), which makes the use of OLS appropriate. Second, OLS is easier to understand by the JMP's non-technical audience, which includes WASH sector stakeholders and policy-makers, and also to implement because results can be generated and displayed without the need of specialized statistical software (JMP, 2014).

4.3. The magnitude of uncertainty around JMP data

In addition to generating "better" WASH estimates by considering both the compositional and non-linear nature of the data, one of the main contributions of our approach is the characterization of uncertainty around estimates. This is done by constructing the 95% confidence intervals around estimates (Table 6).

Confidence intervals are generally wider for the GAM model: in India, for example, the 2020 projection of the percentage of people practicing open defecation (i.e., x_3) is 47.2%–54.8% and 36.0%–61.0% with OLS and GAM, respectively. Furthermore, with few data points, the GAM model results in even wider confidence bounds. For example, in Morocco, the widths of the confidence intervals for 2020 are 3.5, 10.5, 9.6 and 2.2 percentage points (which represent 68.6%, 14.9%, 41.9% and 146.7% of the mean estimated values).

Confidence intervals are generally wider for the GAM model. In India, for example, the 2020 projection of the percentage of people practicing open defecation (x3) is 47.2%–54.8% and 36.0%–61.0% with OLS and GAM, respectively. Furthermore, with few data points, the GAM model results in even wider confidence bounds. For example, in Morocco, the widths of the confidence intervals for 2020 are 3.5, 10.5, 9.6 and 2.2 percentage points (which represent 68.6%, 14.9%, 41.9% and 146.7% of the mean estimated values).

Constructing these confidence intervals around estimates can be extremely useful for two main purposes. First, it allows us to describe the precision of the estimate and represent its sampling distribution. Second, it provides context for policy-making. In the case of rural sanitation in India, 48.7% of population is expected to be practicing open defecation in 2020, with a 95% confidence interval of 36%–61% (Fig. 4). This means that the true coverage of all the population

is likely to be between 36% and 61%, but it might not be: the "95%" indicates that, if we repeated the same household survey many times, 95% of the them would include the true percentage, but 5% would not. Therefore, if the target is to decrease the prevalence of open defecation to below 50% in 2020, we could not be 100% certain that it is achieved, even with a mean point estimate at 48.7%. That is why policy-makers must consider the errors in the WASH coverage estimates when assessing progress against coverage targets. In this sense, the approach we propose can help improve the use of JMP data to evaluate trends in coverage and inform decision making.

However, it is important to recall that our approach only estimates standards errors. As explained in Section 3.1, JMP global database does not provide information on standard errors, because it is rarely included in household survey reports. That is why we approximate standard errors by defining a curve for the relative standard errors, with a fixed maximum and minimum RSE. Clearly, choosing other RSE curves would lead to different confidence intervals (Fig. 5).

Furthermore, it becomes evident the effect of having more precise data. For instance, confidence intervals become considerably narrower around 2014, where the data point comes from a census without sampling errors. Consequently, adding more precise data – either from censuses or household surveys with larger samples – helps reduce the width of the confidence intervals to a point where results are where the results are more reliable and informative. However, larger samples sizes imply higher costs of survey implementation and, more importantly, may introduce more non-sampling errors. Since non-sampling errors can have an inverse relationship with sample size, increasing the later might end up being rather detrimental. That is why the focus should not be only on reducing the level of uncertainty around WASH estimates, but on characterizing and communicating it. Whatever the type and size of the household surveys, estimates of WASH coverage will always have a certain level of error that must be accounted for a better interpretation of coverage trends.

5. Discussion

The use of confidence intervals to characterize uncertainty around regression estimates is not a novelty, and neither is the application of simulation techniques to construct these confidence bounds. Indeed, confidence intervals are widely used in the health sector to communicate uncertainty around child mortality indicators (Bermejo III et al., 2015; Minnery et al., 2015; Hodge et al., 2014). Yet, in the WASH sector, confidence bounds of estimates are rarely reported. The JMP (2014) justifies this by asserting that "it may be more important to be transparent about the level of uncertainty than being able to calculate a quantitative measure that could be misleading". However, how can transparency be realized if margins of error of survey data are not even available in the JMP public database? And even if errors are obtainable, how can they be modelled when the assumption of normally distributed errors is no longer acceptable? These are precisely the two questions we tackle with our approach.

On one hand, our analysis emphasizes the need for recognizing and considering the compositional and non-linear nature of WASH data in order to avoid misleading results. When the compositional parts are analysed separately, the regression models may generate proportion estimates beyond 0% and 100%. Furthermore, when linear regression is applied, WASH coverage can be underestimated or overestimated. Our approach generates more theoretically sound coverage estimates, which could potentially better serve the needs not only of global monitoring agencies such as the JMP but also of country decision-makers.

On the other hand, our uncertainty approach shows that approximating the standard errors with generalized RSE curves solve the issue of non-reporting them. Indeed, this approach can be applied by the international scientific community when dealing with trend analysis of WASH access (Jeuland et al., 2013; Cumming et al., 2014; Beyene et al., 2015; Pérez-Foguet et al., 2017; Armah et al., 2018; Chikozho et al., 2019). Yet, our approach should be viewed merely as an approximation of these errors: in order to obtain results more coherent with reality, more information on standard errors must be provided in household survey reports.

6. Conclusions

Characterizing uncertainty around WASH estimates is crucial when interpreting results, especially when assessing coverage trends over time or comparing coverage across countries. However, reporting confidence intervals in WASH estimates can be challenging since survey data compiled by the JMP does not have publicly available margins of error. In this article, we have presented a simple approach to characterize and communication uncertainty in WASH estimates, and, simultaneously, produce "better" estimates by considering the compositional nature and non-linearity of the data.

Three main conclusions can be drawn:

- WASH data are compositional, and thus should not be modelled with standard statistical analysis. Log-ratio transformations designed for compositional data lead to more conceptually sound estimates, especially in the occurrence of coverage rates near 0% or 100%.
- OLS regression may underestimate or overestimate coverage of WASH services when coverage data show non-linear patterns such as acceleration, deceleration and saturation. Non-linear methods such as GAM may be an alternative to account for the non-linear trajectories in WASH access.
- Standard errors of survey data can be approximated with our approach, but to obtain a more accurate measure of the magnitude of uncertainty around WASH estimates, more efforts should be made to include errors in household survey reports and the JMP global database.

Uncited references

Barde and Barde, 2012
Egozcue and Pawlowsky-Glahn, 2005
LIXIL and Oxford Economics, 2016

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Methodol. 44 (2), 139–160.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, Ltd, London.
Aitchison, J., 1999. Logratios and natural laws in compositional data analysis. Math. Geol. 31 (5), 563–580.
Armah, F.A., Ekumah, B., Yawson, D.O., Odoi, J.O., Afitiri, A.R., Nyieku, F.E., 2018. Access to improved water and sanitation in sub-Saharan Africa in a quarter century. Heliyon 4 (11), e00931.
Bain, L.J., Engelhardt, M., 1991. Introduction to Probability and Mathematical Statistics. Duxbury Press.
Bain, R., Johnston, R., Mitis, F., Chatterley, C., Slaymaker, T., 2018. Establishing sustainable development goal baselines for household drinking water, sanitation and hygiene services. Water 10 (12), 1711.
Banda, J.P., 2003. Nonsampling errors in surveys. In: Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys. United Nations Secretariat, Statistics Division, ESA/STAT/AC.93/7. Available at: https://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_7.pdfAccessed 29 May 2019.

Barde, M.P., Barde, P.J., 2012. What to use to express the variability of data: standard deviation or standard error of mean?. Perspectives in clinical research 3 (3), 113–116.
Bartram, J., Brocklehurst, C., Fisher, M., Luyendijk, R., Hossain, R., Wardlaw, T., Gordon, B., 2014. Global monitoring of water supply and sanitation: history, methods and future challenges. Int. J. Environ. Res. Public Health 11 (8), 8137–8165.
Becker, M., Klößner, S., 2017. PearsonDS: Pearson Distribution System. R package version 1.1. Available at https://cran.r-project.org/web/packages/PearsonDS.
Bergeron-Boucher, M.P., Canudas-Romo, V., Oeppen, J., Vaupel, J.W., 2017. Coherent forecasts of mortality with compositional data analysis. Demogr. Res. 37, 527–566.
Bermejo III, R., Firth, S., Hodge, A., Jimenez-Soto, E., Zeck, W., 2015. Overcoming stagnation in the levels and distribution of child mortality: the case of the Philippines. PLoS One 10 (10), e0139458.
Betti, G., Gagliardi, F., Verma, V., 2018. Simplified Jackknife variance estimates for fuzzy measures of multidimensional poverty. Int. Stat. Rev. 86 (1), 68–86.
Beyene, A., Hailu, T., Faris, K., Kloos, H., 2015. Current state and trends of access to sanitation in Ethiopia and the need to revise indicators to monitor progress in the Post-2015 era. BMC Public Health 15 (1), 451.
Bowman, K.O., Shenton, L.R., 2007. The beta distribution, moment method, Karl Pearson and RA Fisher. Far East Journal of Theoretical Statistics 23 (2), 133.
Cameron, E., 2011. On the estimation of confidence intervals for binomial population proportions in astronomy: the simplicity and superiority of the Bayesian approach. Publ. Astron. Soc. Aust. 28 (2), 128–139.
Chikozho, C., Kadengye, D.T., Wamukoya, M., Orindi, B.O., 2019. Leaving no one behind? Analysis of trends in access to water and sanitation services in the slum areas of Nairobi, 2003–2015. Journal of Water, Sanitation and Hygiene for Development
Cronk, R., Slaymaker, T., Bartram, J., 2015. Monitoring drinking water, sanitation, and hygiene in non-household settings: priorities for policy and practice. Int. J. Hyg. Environ. Health 218 (8), 694–703.
Cumming, O., Elliott, M., Overbo, A., Bartram, J., 2014. Does global progress on sanitation really lag behind water? An analysis of global progress on community-and household-level access to safe water and sanitation. PLoS One 9 (12), e114699.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37 (7), 795–828.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35 (3), 279–300.
Eisele, T.P., Rhoda, D.A., Cutts, F.T., Keating, J., Ren, R., Barros, A.J., Arnold, F., 2013. Measuring coverage in MNCH: total survey error and the interpretation of intervention coverage estimates from household surveys. PLoS Med. 10 (5), e1001386.
Ezbakhe, F., Pérez-Foguet, A., 2019. R Script for Characterizing Uncertainty in Water and Sanitation Estimates, (Available at).
Ferrer-Rosell, B., Coenders, G., Martínez-Garcia, E., 2016. Segmentation by tourist expenditure composition: an approach with compositional data analysis and latent classes. Tour. Anal. 21 (6), 589–602.
Fuller, J.A., Goldstick, J., Bartram, J., Eisenberg, J.N., 2016. Tracking progress towards global drinking water and sanitation targets: a within and among country analysis. Sci. Total Environ. 541, 857–864.
Gabrel, C., Jones, A., 2000. The national nursing home survey: 1995 summary. National Center for Health Statistics. Vital Health Stat. 13 (146), 1–86.
Gerald van den Boogaart, K., Tolosana-Delgado, R., Bren, M., 2018. Compositions: compositional data analysis. R package version 1.40-2. Available at https://cran.r-project.org/web/packages/compositions.
Hastie, T., 2018. Gam: generalized additive models. R package version 1.16. Available at https://cran.r-project.org/web/packages/gam/.
Hodge, A., Firth, S., Marthias, T., Jimenez-Soto, E., 2014. Location matters: trends in inequalities in child mortality in Indonesia. Evidence from repeated cross-sectional surveys. PLoS One 9 (7), e103597.
Janicki, R., 2019. Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. Communications in Statistics-Theory and Methods 1–21.
Jeuland, M.A., Fuente, D.E., Ozdemir, S., Allaire, M.C., Whittington, D., 2013. The long-term dynamics of mortality benefits from improved water and sanitation in less developed countries. PLoS One 8 (10), e74804.
JMP, 2014. Report of WHO/UNICEF JMP Task Force on Methods. Available at WHO/UNICEF https://washdata.org/sites/default/files/documents/reports/2017-07/JMP-Methods-Task-Force-Report-Final.pdf, Accessed 28 May 2019.
JMP, 2015. Progress on sanitation and drinking water: 2015 update and MDG assessment. Available at WHO/UNICEF https://data.unicef.org/wp-content/uploads/2015/12/Progress-on-Sanitation-and-Drinking-Water_234.pdf, Accessed 28 May 2019.
JMP, 2017. Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines. Available at WHO/UNICEF https://www.unicef.org/publications/files/Progress_on_Drinking_Water_Sanitation_and_Hygiene_2017.pdf, Accessed 28 May 2019.
JMP, 2018. JMP methodology: 2017 update and SDG baselines. Available at WHO/UNICEF https://washdata.org/sites/default/files/documents/reports/2018-04/JMP-2017-update-methodology.pdf, Accessed 28 May 2019.
Linares-Mustaros, S., Coenders, G., Vives-Mestres, M., 2018. Financial performance and distress profiles. From classification according to financial ratios to compositional classification. Adv. Account. 40, 1–10.
LIXIL and Oxford Economics, 2016. The True Cost of Poor Sanitation, in Collaboration with WaterAid. Lixil Group Corporation, Tokyo, Japan.
Lloyd, C.D., Pawlowsky-Glahn, V., Egozcue, J.J., 2012. Compositional data analysis in population studies. Ann. Assoc. Am. Geogr. 102 (6), 1251–1266.
Luh, J., Baum, R., Bartram, J., 2013. Equity in water and sanitation: developing an index to measure progressive realization of the human right. Int. J. Hyg. Environ. Health 216 (6), 662–671.
Marcillo-Delgado, J.C., Ortego, M.I., Pérez-Foguet, A., 2019. A compositional approach for modelling SDG7 indicators: case study applied to electricity access. Renew. Sust. Energ. Rev. 107, 388–398.
MICS, 2006. Enquête par Grappes à Indicateurs Multiples (MICS) et Indicateurs des Objectifs du Millénaire pour le Développement (OMD), Burkina Faso. Available at MICS http://mics-surveys-prod.s3.amazonaws.com/MICS3/West%20and%20Central%20Africa/Burkina%20Faso/2006/Final/Burkina%20Faso%202006%20MICS_French.pdf.
Minnery, M., Firth, S., Hodge, A., Jimenez-Soto, E., 2015. Neonatal mortality and inequalities in Bangladesh: differential progress and sub-national developments. Matern. Child Health J. 19 (9), 2038–2047.
Pawlowsky-Glahn, V., Egozcue, J.J., 2006. Compositional data and their analysis: an introduction. Geol. Soc. Lond., Spec. Publ. 264 (1), 1–10.
Pawlowsky-Glahn, V., Egozcue, J.J., 2011. Exploring compositional data with the CoDa-dendrogram. Austrian Journal of Statistics 40 (1&2), 103–113.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons, London.
Pérez-Foguet, A., Giné-Garriga, R., Ortego, M.I., 2017. Compositional data for global monitoring: the case of drinking water and sanitation. Sci. Total Environ. 590, 554–565.
Prüss-Ustün, A., Wolf, J., Bartram, J., Clasen, T., Cumming, O., Freeman, M.C., Gordon, B., Hunter, P.R., Medlicott, K., Johnston, R., 2019. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: an updated analysis with a focus on low- and middle-income countries. Int. J. Hyg. Environ. Health S1438-4639 (18), (31048-4).
UNGA (United Nations General Assembly), 2015. Transforming our world: the 2030 agenda for sustainable development. A/RES/70/1. UN, New York, Available at UNGA https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=EAccessed 28 May 2019.
Vaessen, M., Thiam, M., Lê, T., 2005. Chapter XXII the demographic and health surveys. Available at UNDESA United Nations Statistical Division, United Nations Department of Economic and Social Affairshttp://www.iupui.edu/~histwhs/afrhlth/DHSBackground.pdf, Accessed 10 June 2019.
Verma, V., Lê, T., 1996. An analysis of sampling errors for the demographic and health surveys. International Statistical Review/Revue Internationale de Statistique 265–294.
Wei, Y., Wang, Z., Wang, H., Yao, T., Li, Y., 2018. Promoting inclusive water governance and forecasting the structure of water consumption based on compositional data: a case study of Beijing. Sci. Total Environ. 634, 407–416.
WHS, 2003. WHO world health survey: report of Burkina Faso. Available at WHO https://www.who.int/healthinfo/survey/whsbfa-burkinafaso.pdf.
Wolf, J., Bonjour, S., Prüss-Ustün, A., 2013. An exploration of multilevel modeling for estimating access to drinking-water and sanitation. J. Water Health 11 (1), 64–77.